Sequential Quasi Monte Carlo for Dirichlet Process Mixture Models

Julyan Arbel, Jean-Bernard Salomond

Sequential Quasi Monte Carlo for Dirichlet Process Mixture Models

Julyan Arbel, www.julyanarbel.com, Inria, Mistis, Grenoble, France
Jean-Bernard Salomond, Université Paris-Est, France

Sampling: SMC, QMC, SQMC

- **Sequential Monte Carlo** (SMC), or Particle filtering, is a principled technique which sequentially approximates the full posterior using particles (Doucet et al., 2001). It focuses on sequential state-space models: the density of the observations $\mathbf{y}_t$ conditionally on Markov states $\mathbf{x}_t$ in $\mathcal{X} \subseteq \mathbb{R}^d$ is given by $\mathbf{y}_t|\mathbf{x}_t \sim f^Y(\mathbf{y}_t|\mathbf{x}_t)$, with kernel

$$\mathbf{x}_0 \sim f_0^X(\mathbf{x}_0), \quad \mathbf{x}_t|\mathbf{x}_{t-1} \sim f^X(\mathbf{x}_t|\mathbf{x}_{t-1}). \tag{1}$$

- The initial motivation of **quasi Monte Carlo** (QMC) is to use *low discrepancy* vectors instead of unconstrained random vectors in order to improve the calculation of integrals via Monte Carlo.
- Gerber and Chopin (2015) introduce a **sequential quasi Monte Carlo** (SQMC) methodology. This assumes the existence of transforms Γ_t mapping uniform random variables to the state variables. Requires that (1) can be rewritten as

$$\mathbf{x}_0^{(n)} = \Gamma_0(\mathbf{u}_0^{(n)}) \leftrightarrow \mathbf{x}_0^{(n)} \sim f_0(\mathrm{d}\mathbf{x}_0^{(n)})$$
$$\mathbf{x}_{1:t}^{(n)} = \Gamma_t(\mathbf{x}_{1:t-1}^{(n)}, \mathbf{u}_t^{(n)}) \leftrightarrow \mathbf{x}_{1:t}^{(n)}|\mathbf{x}_{1:t-1}^{(n)} \sim f_t(\mathrm{d}\mathbf{x}_{1:t}^{(n)}|\mathbf{x}_{1:t-1}^{(n)})$$

where $\mathbf{u}_t^{(n)} \sim \mathcal{U}([0,1)^d)$ is to be a quasi random vector of uniforms.

Dirichlet process & SQMC

- Nonparametric mixtures for density estimation: extension of finite mixture models when the number of clusters is unknown. Observations $\mathbf{y}_{1:T}$ follow a DPM model with kernel ψ parameterized by $\theta \in \Theta$,

$$\mathbf{y}_t|G \overset{\text{i.i.d.}}{\sim} \int \psi(y;\theta)\mathrm{d}G(\theta), \quad t \in (1:T),$$

where $G \sim \mathrm{DP}(\alpha, G_0)$.

- DPM cast as SMC samplers by Liu (1996); Fearnhead (2004); Griffin (2015) : observations are spread out into unobserved clusters whose labels, or allocation variables, are *latent* variables acting as observations *states* in the context of SMC. Transition is given by the (posterior) *generalized Pólya urn scheme*

$$p_{t,j} = \mathrm{P}(\mathbf{x}_t = j|\mathbf{x}_{1:t-1}, \mathbf{y}_{1:t}).$$

- Complies with Gerber and Chopin (2015) need for a deterministic transform

$$\Gamma_t(\mathbf{x}_{1:t-1}^{(n)}, \mathbf{u}_t^{(n)}) =$$
$$\min\left\{j \in \{1,\dots,k_{t-1}^{(n)}+1\} : \sum_{i=1}^{j} p_{t,i}^{(n)} > \mathbf{u}_t^{(n)}\right\}$$

for any particle n, with $\mathbf{u}_t^{(n)} \sim \mathcal{U}([0,1))$.

Goal

- Peculiarity to the DPM setting:
  - state-space $\approx (1:T)^T$ is discrete and varies
  - transition is not Markovian
- **Goal**: investigate how SQMC fares
  - compare allocation trajectories $\mathbf{x}_{1:T}^{(n)}$, $n = 1,\dots,N$ in SMC & SQMC
  - measure their dispersion with a principal component analysis (PCA) → proportion of variance explained by number of components in the PCA

References

Doucet, A., de Freitas, N., and Gordon, N. J. (2001). *Sequential Monte Carlo methods in Practice*. Springer-Verlag, New York.

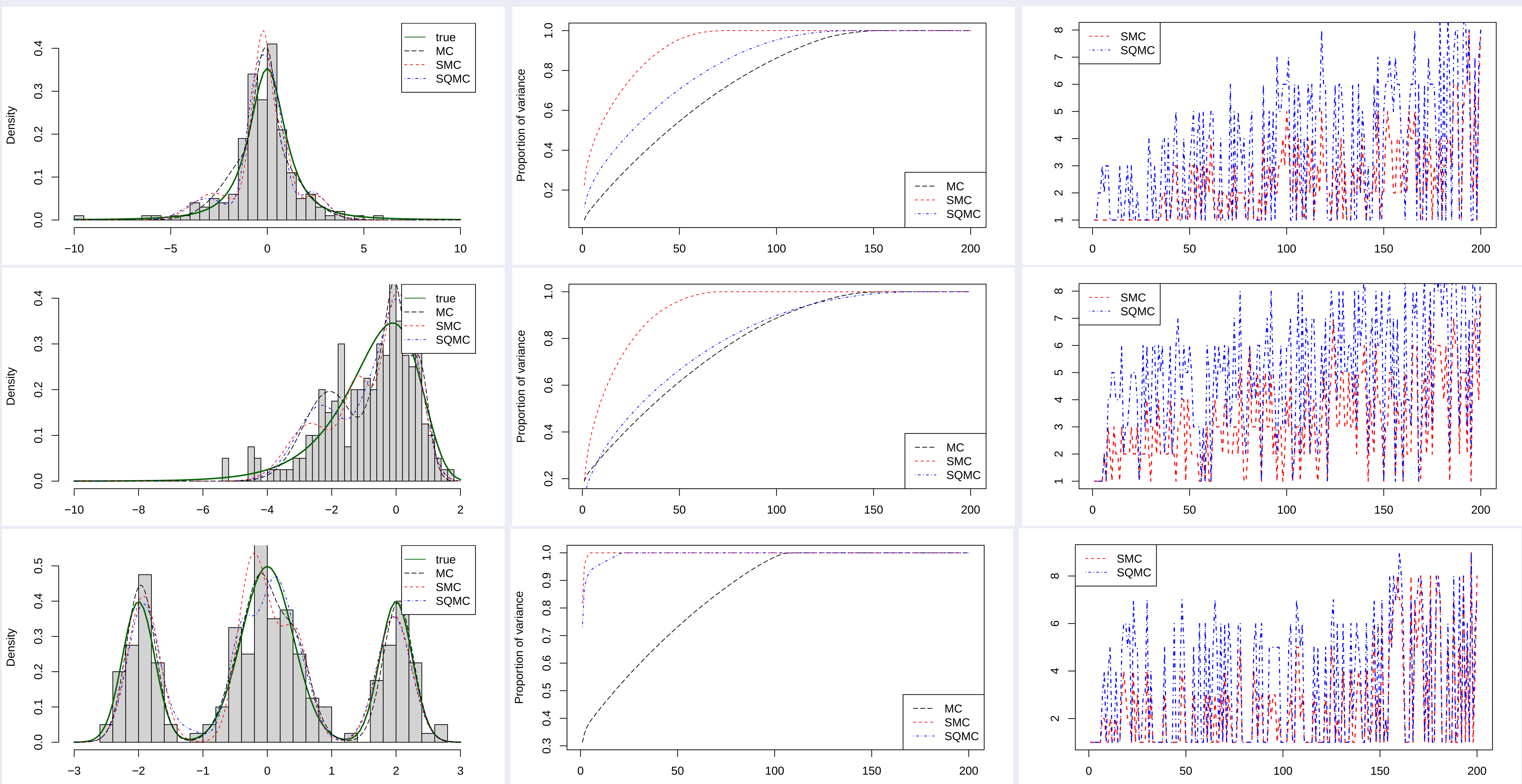
Fearnhead, P. (2004). Particle filters for mixture models with an unknown number of components. *Statistics and Computing*, 14(1):11–21.

Gerber, M. and Chopin, N. (2015). Sequential quasi Monte Carlo. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(3):509–579.

Griffin, J. E. (2015). Sequential Monte Carlo methods for mixtures with normalized random measures with independent increments priors. *Statistics and Computing, in press*.

Liu, J. S. (1996). Nonparametric hierarchical Bayes via sequential imputation. *The Annals of Statistics*, 24:910–930.

Results II



Left: **Density fit**; Middle: **Particles diversity** (PCA); Right: **Number of different particles for each data point**.
Three samplers, non sequential Monte Carlo, **MC**, sequential Monte Carlo, **SMC** and sequential quasi Monte Carlo **SQMC**.
Sample size $T = 200$, number of particles $N = 1\,000$.
Top row: Heavy tailed distr. (student 2); Middle row: Skewed distr. (log-Gamma); Bottom row: Multimodal distr. (mixture of normals).